Chapter 7

End-to-End Routing Stability

One key property we would like to know about an end-to-end Internet route is its *stability*: do routes change often, or are they stable over time? In this section we analyze the routing measurements to address this question. We begin by discussing the impact of routing stability on different aspects of networking, to motivate our study, and summarizing the reasons why routes change. We then present two different notions of routing stability, "prevalence" and "persistence," and show that they can be orthogonal (i.e., a route can be considered "stable" by one definition independently of whether it is stable by the other definition).

It turns out that "prevalence" is quite easy to assess from our measurements, and "persistence" quite difficult. In § 7.5 we characterize the "prevalence" stability of the routes, and then in § 7.6 we tackle the problem of assessing "persistence."

We finish by evaluating a method for *detecting* route changes based on observing changes in hop count (TTL). We find this method makes a decent heuristic, but generates enough "false negatives" that it should not be trusted if accuracy is crucial.

7.1 Importance of routing stability

One of the stated goals of the Internet architecture is that large-scale routing changes (i.e., those involving different autonomous systems) rarely occur [Li89]:

> The Inter-AS Routing scheme must provide stability of routes. It is totally unacceptable for routes to vary on a frequent basis. This requirement is not meant to limit the ability of the routing algorithm to react rapidly to major topological changes, such as the loss of connectivity between two AS's. The need for adaptive routing does not imply any desire for load-based routing.

This point has been argued by others as well [BE90, Tr95b]. Routing instability sets the foremost limit on how use of BGP can scale to a very large internet, because CPU utilization required by BGP routers increases directly in proportion to the frequency of routing changes (but not, otherwise, in proportion to the overall size of the network) [Tr95b]. Hence, the key concern is that routing instability can in turn lead to general network instability (i.e., loss of packet-forwarding function).

There are a number of aspects of networking affected by routing stability:

1. Some of the most important properties of a network—latency, bandwidth, congestion levels, packet losses—are all *route* properties. If the route through the network changes, so might some or all of these properties. Therefore, the degree to which a network's behavior is *predictable* is directly related to the stability of its routes. This is not to say that, even if the route remains stable, these properties will too. Rather, routing stability is *necessary* but not *sufficient* for predictable network behavior.

   One particular example affected by routing stability is the *predictive service* scheme proposed for real-time network traffic [CSZ92]. Predictive service attempts to satisfy the performance requirements of real-time traffic by only admitting new real-time flows if recent traffic measurements suggest the network has sufficient capacity for them. If routes are unstable over short time scales, however, then these predictions become considerably difficult to make.

2. The degree to which endpoints can benefit from *caching* information of previously encountered path conditions is limited by (among other factors) whether the route observed in the past is likely to be the same as the present route.

3. New network protocols supporting "real-time" applications such as audio and visual flows generally require establishing state in routers in order to assure that the flows receive the necessary performance. Real-time flows will often be long-lived, existing for time spans on the order of human interactions (minutes to hours) rather than computer interactions (milliseconds to seconds). If routing changes occur frequently, then these long-lived flows will be prone to losing the state they have established in the routers in the network, and will suffer outages or degraded service while they attempt to find alternate routes with sufficient resources.

   Some protocols use "hard state" in the routers, meaning that, if state information for a given flow is not present in the router, then the router will not forward the flow's packets [DB95, FBZ94]. Other protocols use "soft state" schemes in which, even if a router has no corresponding state information for the flow, it will forward a flow's packets, though with possibly degraded performance [ZDESZ93, BCS94, DEFJLW94]. Hard state and soft state schemes trade off performance guarantees versus flexibility in the face of errors. Part of the question of evaluating the flexibility gain of soft state schemes concerns the degree of route stability. If routes do not tend to change frequently, then the soft state gain in flexibility is minor, but, if routes change frequently, then the gain will be larger.

   For an overview of the difficulties of dealing with routing changes in real-time protocols, see [GR95]. We do not attempt here to evaluate the flexibility gain of soft state versus hard state schemes. Indeed, the question is much more complex than stated above[1]. But we do attempt to characterize the stability constants that could then be used in such an evaluation.

4. Another form of router state arises from schemes for supporting *advance reservations*, in which the network allows resources to be reserved for future use [FGV95]. If the state con-

[1] For example, both types of schemes often use "route pinning," in which the route available when a flow is established remains the route used by that flow for its lifetime. If a route is pinned, then only route changes due to the *failure* of a router used by the flow affect the flow; not those due to the discovery of improved routes (§ 7.2).

Similarly, some hard state schemes have explicit recovery mechanisms for when a flow's route *does* fail ([Ba94, DB95, GR95]), so these schemes do not necessarily stop working in the presence of route changes.

cerning these reservations is stored in the network's routers (a logical choice, to avoid centralized bottlenecks), then frequent route changes may lead to reservations failing because the routers used to establish the reservations are no longer the routers relevant to the real-time path.

5. If routes change frequently, then network measurements face difficult consistency problems. For example, several studies of end-to-end network behavior rely on repeated measurements of a network path made over the course of hours to days [Mi83, CPB93a, Bo93, SAGJ93, Mu94, BCG95]. Whether these measurements all observe the same path significantly affects the accuracy of the studies.

   Similarly, distributed algorithms for analyzing the network's state also face consistency problems if routes change frequently. For example, recent theoretical work has developed "tomography" techniques for inferring end-to-end network traffic intensities using just measurements of aggregate traffic intensities along the network's links [Va95]. The work assumes stable routing (an extension explores Markovian routing). If routes change frequently, then it may prove extremely difficult to capture a consistent global snapshot of any significant portion of the Internet for purposes of operational monitoring.

We now look briefly at why routes change, and then introduce two different notions of routing stability, to encompass the different stability concerns discussed above.

7.2 Why routes change

There are several different reasons why a route might change:

1. If a link or router *fails*, then the network must reroute traffic using that link or router.

2. If a link or router *recovers*, then the network *may* elect to route previously redirected traffic back to using that link or router. If routes are "pinned," however, then they will not be changed due to recoveries.

3. If a link *degrades* or *improves*, where such notions might for example be measured by congestion levels, then the network might *adapt* by changing routes to account for the altered view of the cost of the link. For example, the ARPANET routing algorithms were designed to route around congested areas of the network. As experience with the ARPANET showed, such adaptive routing is tricky to get right: the initial routing scheme reacted "very quickly to good news, and very slowly to bad news" [MFR78], and the first revision of the algorithm [MRR80] also exhibited oscillations under heavy load [KZ89]. Because it is difficult to achieve stable adaptive routing, in which routes are not subject to rapid oscillation in response to transient congestion, adaptive routing is not widely used [Mo95], and a number of researchers argue for caution in its use [ERH92, RG95].

4. A router might cycle between different routes to the same destination in order to *balance load*. We analyzed this sort of route "flutter" in § 6.6, where we found that often its effects are confined to a single hop in an Internet path, but sometimes the split routes fail to rejoin, leading to drastically different path characteristics.

We would hope to observe four different time constants associated with these four reasons, of decreasing durations. Link failures should occur only rarely, hopefully on the time scale of days. Link recoveries should occur significantly quicker (i.e., shortly after the link failure), on the time scale of minutes (if a reboot or restart is all that is required) to hours (if human intervention and repair is required). If adaptive routing is used, then changes should occur on the time scales of congestion epochs (unfortunately not well characterized in the literature), which one presumes is on the order of seconds to minutes; adaptive routing algorithms generally *damp* rapid changes, though, to avoid oscillations, so we would expect this time constant to be more on the order of minutes. Finally, load balancing is generally done on very small time scales (such as every other packet), on the order of milliseconds.

7.3 Two definitions of stability

As suggested in § 7.1, there are two distinct views of routing stability. The first is: "Given that I observed route r at time t, how likely am I to observe r again at time $t+s$?" We refer to this notion as *prevalence*. A route's prevalence directly affects the first two motivations discussed above, namely predictability of service, and our ability to learn from past conditions. In general, the degree of route prevalence will depend on s. For large s, however, we would expect the observation at time $t+s$ to be (nearly) independent of the observation at time t. In this study, for simplicity we focus on the unconditional probability of observing a route, confining our analysis to $s \to \infty$, i.e., the steady-state probability of observing r again at a point far in the future. We leave the interesting question of how prevalence evolves for different intervals s for future work.

A second view of stability is: "Given that I observed route r at time t, how long before that route is likely to have changed?" The likelihood of routes changing in the near future has implications for the latter three motivations, namely hard and soft router state, resource reservations, and network measurement consistency. We refer to this notion as *persistence*.

Intuitively, we might expect these two notions to be coupled. Consider, for example, a sequence of routing observations made every T units of time. If the routes we observe are:

$$R_1, R_1, R_1, R_1, R_1, R_1, R_1, R_1, R_1, R_1, R_1, R_2, R_1, R_1, R_1 \ldots$$

then clearly route R_1 is much more prevalent than route R_2. We might also conclude that route R_1 is persistent, because we observe it so frequently; but this is not at all necessarily the case. For example, suppose T is one day. If the mean duration of R_1 is actually 10 days, and that of R_2 is one day, then this sequence of observations is quite plausible, and we would be correct in concluding that R_1 is *persistent and prevalent*. Furthermore, depending on our concern, we might also deem that R_2 is persistent, since on average it lasts for a full day (if its lifetime were much shorter, then we would have been unlikely to observe it from measurements made only once a day). If we consider a route that last for more than a few hours as persistent, then from the above observations we could argue that R_2 is *persistent but not prevalent.*

But suppose instead that the mean duration of R_1 is 10 seconds and the mean duration of R_2 is 1 second, and that alterations between them occur as a semi-Markov process,[2] where state 1

[2]Such processes consist of a set of states. Each state i has associated with it a distribution of durations, G_i. The distribution depends on the state number i, but not on anything else. Upon entering state i, a duration is drawn independently

of the process corresponds to R_1, state 2 to R_2, and $P_{1,2} = P_{2,1} = 1$ (i.e., whenever a change occurs, it is a change to the other route). Then a well-known result from the theory of stochastic processes states that the proportion of time the system spends in state 1 is equal to the mean duration of state 1 divided by the sum of the mean durations of states 1 and 2 [Ro83]. For our example, we have that the proportion of time spent in state R_1 is $\frac{10}{11}$, reflecting that R_1 is prevalent. Similarly, the proportion of time spent in state R_2 is $\frac{1}{11}$. Given these proportions, the sequence of observations is *again plausible*, even though each observation of R_1 is actually of a separate instance of the route. In this case, R_1 is *prevalent but not persistent*, and R_2 is *neither prevalent nor persistent*. In other words, we very likely are missing instances of R_2 between observations of R_1, and hence R_1 is not persistent.

This example shows that the notions of "prevalent" versus "persistent" stability are orthogonal, in the sense that the presence or absence of one does not necessarily indicate anything about the presence or absence of the other.

7.4 Reducing the data

To begin our analysis, we first need to reduce the more than 40,000 `traceroutes` measurements in $\mathcal{R}_1$ and $\mathcal{R}_2$ to those relevant for assessing stability. Before we had gathered the $\mathcal{R}_2$ measurements, we performed an initial stability analysis of the $\mathcal{R}_1$ data. Doing so, we concluded that the inter-measurement spacing of the $\mathcal{R}_1$ `traceroutes` , on average about one day, was too large to allow any assessment of routing stability in terms of persistence, because of the ambiguities discussed in the previous section. Consequently, we confine our routing stability analysis to $\mathcal{R}_2$, which contains the bulk (85%) of the 40,000 measurements. 60% of these were taken with a 2-hour inter-measurement spacing. As shown in the remainder of this chapter, this granularity is sufficient to resolve the persistence ambiguities.

Of the 35,109 $\mathcal{R}_2$ measurements, we began by excluding those exhibiting the pathologies discussed in Chapter 6, because they reflect connectivity difficulties distinct from routing instabilities.[3] (We did not exclude "circuitous" routes, however, because, as mentioned in § 6.9, these are not true pathologies.) Doing so eliminated 805 `traceroutes`.

We also omitted `traceroutes` for which one or more hops were completely missing (all three of the probe packets unanswered). These measurements are inherently *ambiguous*, because we could not tell if the route was the same as that observed at other instances. This decision eliminated another 2,595 measurements, leaving us with a total of 31,709 measurements.

We next made a preliminary assessment of the patterns of route changes by seeing which changes occurred the most frequently. We found the pattern of changes dominated by a number of

---

from G_i. The process remains in state i until the duration elapses. At this point, a new state j is chosen based on a set of probabilities fixed for state i.

[3] An exception is the pathology of a routing change during a `traceroute`. Including these pathologies, however, can lead to overestimating the frequency of route changes. Suppose we make three route measurements of a particular path, yielding routes A, A/B, and B, where A/B indicates a `traceroute` that included a change from route A to route B. If we included the second, pathological measurement, we would conclude that over the three observations two changes occurred (A to A/B and A/B to B), whereas in reality only one change occurred (A to B).

It is possible that instead the sequence we observe is A, A/B, A, because route B was short-lived; in this case, omitting the pathological `traceroute` underestimates the frequency of changes. But this becomes an issue only if B was quite short-lived, and we account for such routes separately, as discussed in § 7.6.1.

Routers	Notes
asd01.nl.net, amf01.nl.net	These routers are located in different cities, but provide equal bandwidth and latency to their peers [Lin96].
icm-dc-1.icp.net, icm-dc-2b-s4/0-1984k.icp.net	
rgnet-b1-serial2-3.seattle.mci.net, rainnet-inc.seattle.mci.net	
rb1.rtr.unimelb.edu.au, rb2.rtr.unimelb.edu.au	
unit-gw.unit.no, sintef-gw.sintef.no	Both at the University of Trondheim.

Table XI: Tightly-coupled routers

single-hop differences, at which consecutive measurements showed exactly the same path except for a single router. Furthermore, the names of these routers often suggested that the pair were administratively interchangeable.[4] For example, many of the routing changes to the `austr` site only differed in whether the University of Melbourne border router in the route was `rb1.rtr.unimelb.edu.au` or `rb2.rtr.unimelb.edu.au`. Which of these two routers provides the route to the `austr` host depends on the distribution of load within other parts of the University, but the two routers are under the same direct administration and would indeed be one machine if a single router with sufficient capacity had been available at the time of acquisition [El96].

It seems likely that many route changes differing at just a single hop are due to shifting traffic between two tightly coupled machines. For the stability concerns given in § 7.1, such a change is likely to have little consequence, provided the two routers are co-located and capable of sharing state. We decided that, if a single pair of routers with like names were responsible for more than 200 routing transitions, then we would classify them as "tightly coupled," and merge them into a single router for purposes of evaluating stability. Table XI summarizes these routers. After merging those responsible for > 200 changes, the remaining pairs were all responsible for 80 or fewer changes. We left these as separate routers, as changes between them did not dominate the data, and we would like to minimize assumptions about which routers are tightly coupled.

Finally, we reduced the acceptable routes to three different levels of *granularity*. First, we considered each route as a sequence of Internet hostnames. We call this *host* granularity. We then reduced the routes to sequences of *cities*, as outlined in § 5.3. Note that a route change at host granularity might *not* be a route change at *city* granularity, though the converse always holds. The motivation behind the distinction of host granularity vs. city granularity is to introduce a notion of "any change" vs. "major change." A route change at city granularity will likely have considerably more repercussions than a change visible only at host granularity. For example, the latency of the route will often be different. Overall, 57% of the route changes at host granularity were also route changes at city granularity.

[4] Sometimes the routers *were* identical. For example, IP address `192.157.65.130`, which translates to `icm-paris-1-s0-1984k.icp.net`, is actually also an interface on `paris-ebs2.ebone.net`.

The third level of granularity was *AS path*—the sequence of autonomous systems visited by the route (§ 4.4). A change at *AS* granularity reflects a possible change in the intermediate routing algorithms and policies, and as such is another form of major change. Overall, 36% of the route changes at host granularity were also changes at AS path granularity. Note that a change at AS path granularity is not necessarily a change at city granularity, nor vice versa, though overall we found AS path granularity coarser (i.e., comprising fewer changes) than city granularity.

7.5 Routing Prevalence

In this section we look at routing stability from the standpoint of *prevalence*: how likely we are, overall, to observe a particular route (c.f. § 7.3). We can associate with prevalence a parameter π_r, the steady-state probability that a path at an arbitrary point in time uses a particular route r.

We can assess π_r from our data as follows. We suppose that routing changes follow a semi-Markov process. In this model, each route's duration has a fixed distribution (but different routes can have different distributions), and the duration of each instance of a route is independent of all previous route durations. Furthermore, the probability that route r_1 is followed by route r_2 is fixed and independent of past events.

We then use the result that, for a semi-Markov process, the steady-state probability of observing a particular state is equal to the average amount of time spent in that state [Ro83].[5] Furthermore, because of PASTA, our independent exponential sampling gives us an unbiased estimator of this time average (§ 4.3). Suppose we make n observations of a path and k_r of them find state r (i.e., route r). Then we will estimate π_r as $\hat{\pi}_r = k_r/n$.

We proceed as follows. For a particular path p (and for a given granularity), let n_p be the total number of `traceroutes` measuring that path, and d_p the number of distinct routes seen. We will denote the most commonly occurring route as the *dominant* route, and others as *secondary* routes. Thus, there are always $d_p - 1$ secondary routes. Let k_p be the number of times we observe the dominant route. We then confine our analysis to:

$$\hat{\pi}_{\text{dom p}} = k_p/n_p,$$

the prevalence of the dominant route.

Figure 7.1 shows the cumulative distribution of the prevalence of the dominant routes over all of the paths in our study (i.e., all 1,054 source/destination pairs), for the three different granularities. For example, at host granularity, nearly half (49%) of the paths (y-axis) were dominated by a route with a prevalence of at least 80% (x-axis).

There is clearly a wide range, particularly for host granularity. For example, for the path between `pubnix` and `austr`, in 46 measurements we observed 9 distinct routes at host granularity, and the dominant route was observed only 10 times, leading to $\hat{\pi}_{\text{dom}} = 0.217$. On the other hand, at host granularity more than 25% of the paths exhibited only a single route ($\hat{\pi}_{\text{dom}} = 1$). For city and AS path granularities, the spread in $\hat{\pi}_{\text{dom}}$ is more narrow, as would be expected (the figure also

[5] This result requires that the distribution of time spent in each state be *nonlattice*: i.e., not always an integral multiple of some constant, so that the notion of "steady state" can be defined without reference to specifics about exactly when, in the far future, we observe the process. For route durations, this seems like a plausible assumption.

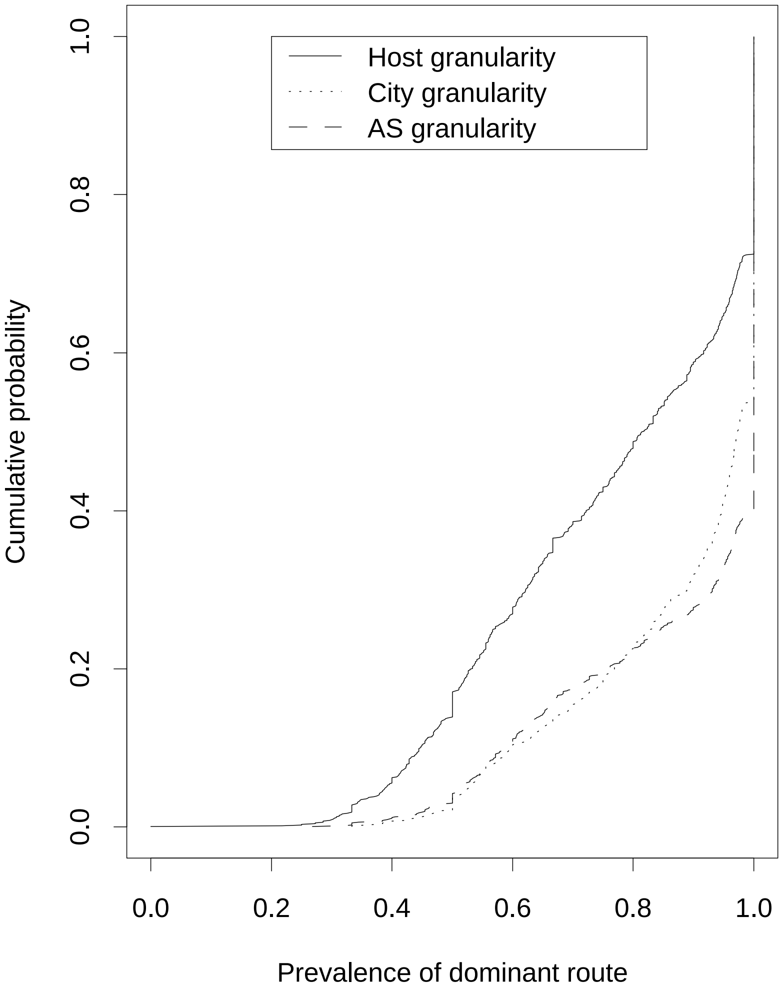

Figure 7.1: Fraction of measurements observing the dominant route, for all paths, at all granularities

shows how route changes at city or AS path granularity do not necessarily imply changes at the other granularity, since neither is strictly below the other).

A key figure to keep in mind from this plot, however, is that, while there is a wide range in the distribution of $\hat{\pi}_{\text{dom}}$ over different paths, its *median* value at host granularity is 82%; 97% at city granularity, and 100% at AS path granularity. The clustering of many paths only ever exhibiting a single route (i.e., prevalence = 100%) reflects the finding we develop below in § 7.6 that many routes are long-lived. (If we had data gathered over periods of time exceeding several weeks, we would doubtless find that the spike at prevalence = 100% would spread out to values in the upper 90%'s.) Thus, we can conclude: *In general, Internet paths are strongly dominated by a single route.*

Our previous work, however, has shown that many characteristics of network traffic exhibit considerable site-to-site variation [Pa94a], and thus it behooves us to assess the differences in $\hat{\pi}_{\text{dom}}$ between the sites in our study. To do so, for each site s (and for each granularity) we computed:

$$\hat{\pi}_{\text{src s}} = \frac{\sum_{\text{src paths } s_i} k_{s_i}}{\sum_{\text{src paths } s_j} n_{s_j}}.$$

where k_{s_i} is the number of times we observed the dominate route when measuring a path from source s to destination i, and n_{s_j} is the total number of times we made a measurement of the path from source s to destination j.

The aggregate estimate $\hat{\pi}_{\text{src s}}$ then indicates the overall prevalence of dominant routes from s to different destinations. We expect variations in this estimate for different sites to reflect differing routing prevalence due to route changes *near* the source. Route changes further downstream from the source occur either deep inside the network (and so will affect many different sites), or near the destination (and thus will not affect any particular *source* site unduly).

Similarly, we can construct $\hat{\pi}_{\text{dst s}}$ for all of the paths with destination s. Studying $\hat{\pi}_{\text{src s}}$ and $\hat{\pi}_{\text{dst s}}$ for different sites and at different granularities reveals considerable site-to-site variation, in agreement with the general findings in [Pa94a]. Figure 7.2 shows the values computed for $\hat{\pi}_{\text{src s}}$ for each of the $\mathcal{R}_2$ sites, at host granularity. We find that the prevalence of the dominant routes originating at the `ucl` source is under 50% (we will see in § 7.6.1 the main cause for this), and for `bnl`, `sintef1`, `sintef2`, and `pubnix` is around 60%; while for `ncar`, `ucol`, and `unij`, it is just under 90%. Even at AS path granularity, the `ucl` source has an average prevalence of 60%, with `ukc` about 70%, and the remainder from 85% to 99%. At city granularity, however, the main outlier is `bnl`, with a prevalence of 75% (c.f. § 7.6.2), because the `ucl` and `ukc` instabilities, while spanning autonomous systems, do not span different cities.

We find similar spreads for $\hat{\pi}_{\text{dst s}}$ for different destination sites s. Figure 7.3 shows the per-site values, computed for host granularity. Sometimes the sites with low overall prevalence are the same as the sites with low prevalence for $\hat{\pi}_{\text{src s}}$ (e.g., `ucl`), and sometimes they are different (e.g., `ukc`); this variation is due to *asymmetric* routing, which we analyze in Chapter 8.

We can thus summarize routing prevalence as follows: *In general, Internet paths are strongly dominated by a single route, but, as with many aspects of Internet behavior, we also find significant site-to-site variation.*

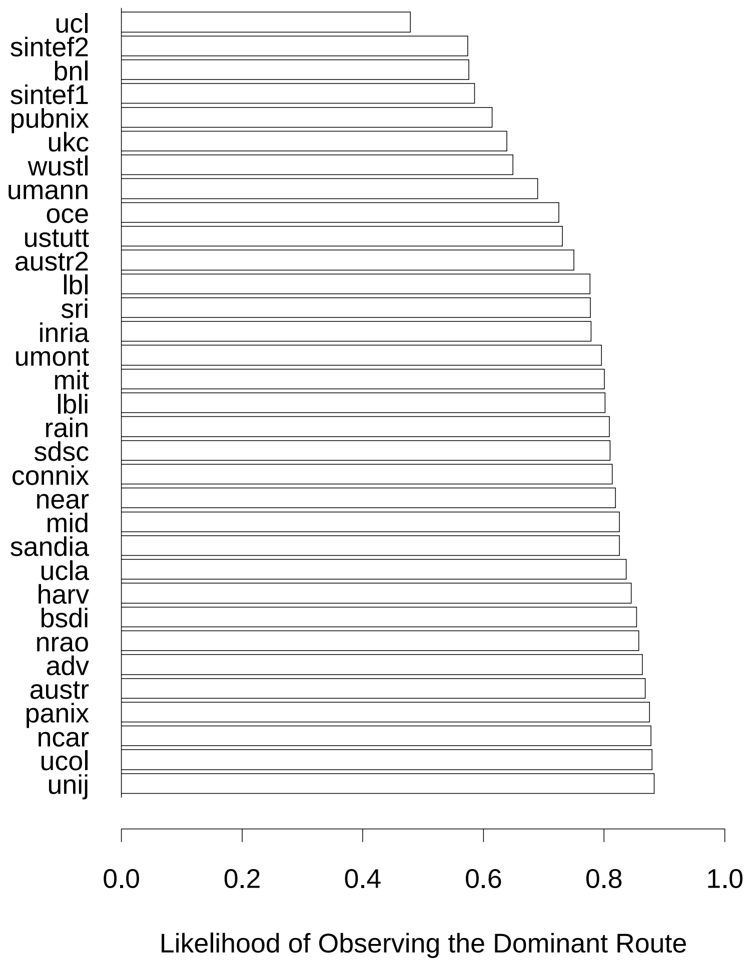

Figure 7.2: Fraction of measurements observing the dominant route, for different source sites, at host granularity

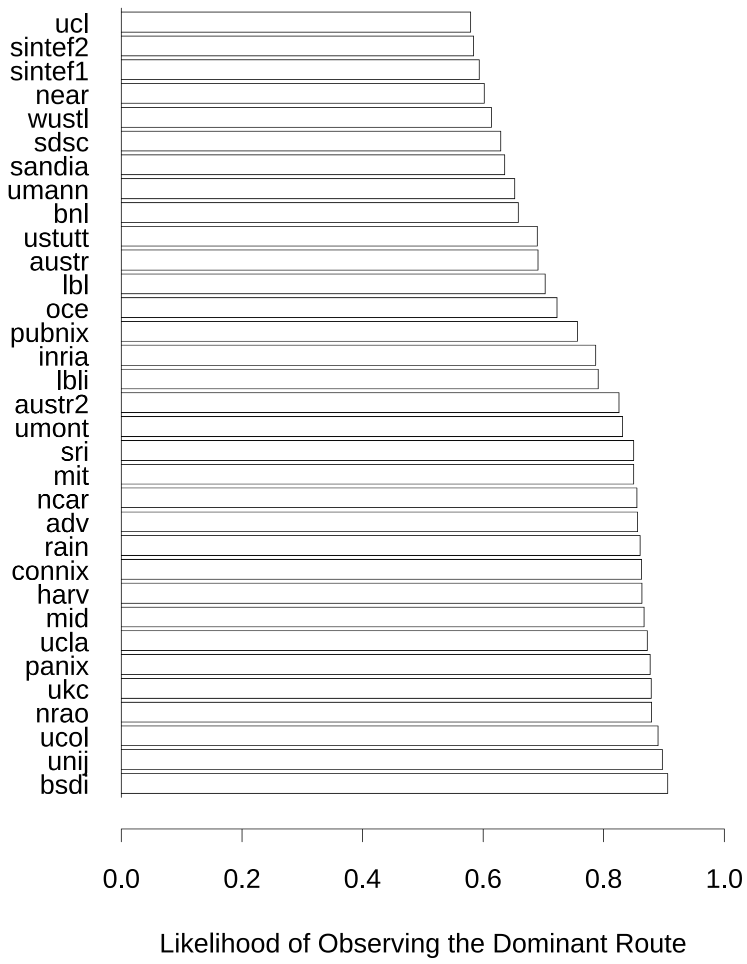

Figure 7.3: Fraction of measurements observing the dominant route, for different destination sites, at host granularity

7.6 Routing Persistence

We now turn to the more difficult task of assessing the *persistence* of routes: How long they are likely to endure before changing. As illustrated in § 7.3, unlike *prevalence*, routing persistence can be difficult to evaluate because a series of measurements at particular points in time do not necessarily indicate a lack of change *and then change back* in between the measurement points. Thus, to accurately assess persistence requires first determining whether routing alternates on short time scales. If not, then we can trust shortly spaced measurements observing the same route as indicating that the route did indeed persist during the interval between the measurements. If shortly spaced measurements can be trusted in this fashion, then they can be used to assess whether routing alternates on medium time scales.

Fortunately, we have measurements made at a number of different intervals: about 60% of the $\mathcal{R}_2$ measurements were exponentially distributed with a mean of 2 hours, and the other 40% with a mean of about 66 hours (with wide variation in the actual intervals, since they were exponentially distributed). While these measurements do not allow us to directly address the problem of assessing persistence—doing so would require a way to unambiguously determine exactly when a route changed, which could be done by tracing BGP routing information exchanges,[6] but not from end-to-end `traceroutes`—our strategy is to analyze the measurements with the shorter spacing to assess the frequency of route alternations, and, in turn, to determine to what degree we can trust the measurements with larger spacing. In this fashion, we aim to "bootstrap" ourselves into a position to be able to make sound characterizations of routing persistence across a number of time scales.

7.6.1 Rapid route alternation

In order to reliably analyze widely-spaced `traceroute` measurements, we must first assess the predominance of rapidly alternating routes. We have already identified two types of rapidly alternating routes, those due to "flutter" and those due to "tightly coupled" routers. We have separately characterized fluttering (§ 6.6) and consequently have not included paths experiencing flutter in this analysis. As mentioned in § 7.4, we merged tightly coupled routers into a single entity, so their presence also does not further affect our analysis of rapidly alternating routes.

We next note that in $\mathcal{R}_2$ we observed 155 instances of a route change during a `traceroute`. The combined amount of time observed by the 35,109 $\mathcal{R}_2$ `traceroutes` was 881,578 seconds. (That is, the mean duration of a $\mathcal{R}_2$ `traceroute` was 25.1 seconds.) Since when observing the network for 881,578 seconds we saw 155 route changes, we can estimate that on average we will see a route change every 5,687 seconds ($\approx$ 1.5 hours). This reflects quite a high rate of route alternation, and bodes ill for relying on measurements made much more than a few hours apart (though see § 7.6.2); but it is not such a high rate that we would expect to completely miss routing changes for sampling intervals significantly less than an hour.

We first looked at those `traceroute` measurements that were made less than 60 seconds apart. There were only 54 of these, but all of them were of the form "R_1, R_1"—i.e., both of the measurements observed the same route. This provides credible, though not definitive, evidence that

[6] As briefly mentioned in § 3.2, recent work by Jahanian, Labovitz and Malan pursues this approach with very interesting results [JLM97]. We became aware of this work too late to discuss it here, but will address it in the version of [Pa96b] that we are presently revising for publication in *IEEE/ACM Transactions on Networking*.

there are no additional widespread, high-frequency routing oscillations, other than those we have already characterized.

We then looked at measurements made less than 10 minutes apart. There were 1,302 of these, and 40 *triple* observations (three observations all within a ten minute interval). The triple observations allow us to double check for the presence of high-frequency oscillations: if we observe the pattern R_1, R_2, R_1 or R_1, R_2, R_3, then we are likely to miss some route changes when using only two measurements 10 minutes apart. If we only observe R_1, R_1, R_1; R_1, R_2, R_2; or R_1, R_1, R_2, then measurements made 10 minutes apart are not missing short-lived routes. Of the 40 triple observations, none were of the form R_1, R_2, R_1 or R_1, R_2, R_3, confirming the finding from the 60 second observations that there are no additional sources of high-frequency oscillation.

The 1,302 ten-minute observations included 25 instances of a route change (R_1, R_2). This suggests that the likelihood of observing a route change over a ten minute interval is not negligible, and requires further investigation before we can look at more widely spaced measurements.

A natural question to ask concerning 10-minute changes is whether they are equally likely to occur along paths between any two sites, or if just a few sites are responsible for most of the 10-minute changes.[7] This is an important consideration: if all paths are equally likely to exhibit a change during a 10-minute interval, then from the figure above of 25 changes observed out of 1,302 ten-minute observations we could conclude that routes change, on average, 25 times per $(1,302 \cdot 10$ min), or about once every eight hours.

We test whether paths to or from particular sites are more prone to change than others as follows. For each site s, let $N^{10}_{\text{src s}}$ be the number of 10-minute pairs of measurements originating at s, and $X^{10}_{\text{src s}}$ be the number of times those pairs reflected a *transition* (i.e., the pair was R_1, R_2). Similarly, define $N^{10}_{\text{dst s}}$ and $X^{10}_{\text{dst s}}$ for those pairs of measurements with destination s. Here we are aggregating, for each site, all of the measurements made using that site as a source (destination), in an attempt to see whether route oscillations are significantly more prevalent near a handful of the sites.

For each site s, we can then define:

$$P^{10}_{\text{src s}} = \frac{X^{10}_{\text{src s}}}{N^{10}_{\text{src s}}},$$

and similarly for $P^{10}_{\text{dst s}}$. These values then give the estimated probability that a pair of ten-minute observations of paths with source (or destination) s will show a routing change. We now check the $P^{10}_{\text{src s}}$ (and $P^{10}_{\text{dst s}}$) estimates for each site to determine which sites appear particularly prone to exhibiting changes during ten minute intervals.

Figure 7.4 shows the sorted $P^{10}_{\text{dst s}}$ estimates. We see, for example, that none of the 10-minute measurements of paths to the destination `adv` observed a route change, but more than 12% of those to `austr` did. From the plot, `austr` appears to be an outlier, and merits further investigation. Before removing it as an outlier, however, we must be careful to first look at its routing oscillations to see what patterns they exhibit.

For the destination `austr`, the 10-minute changes involve a number of source sites: `inria`, `mit`, `near` (twice), and `pubnix`. All of the changes take place at the point-of-entry

[7] Certainly no single path (between the same source/destination pair) is skewing the count of 10-minute changes, since the most frequently observed single path only accounted for 8 of the 1,302 observations.

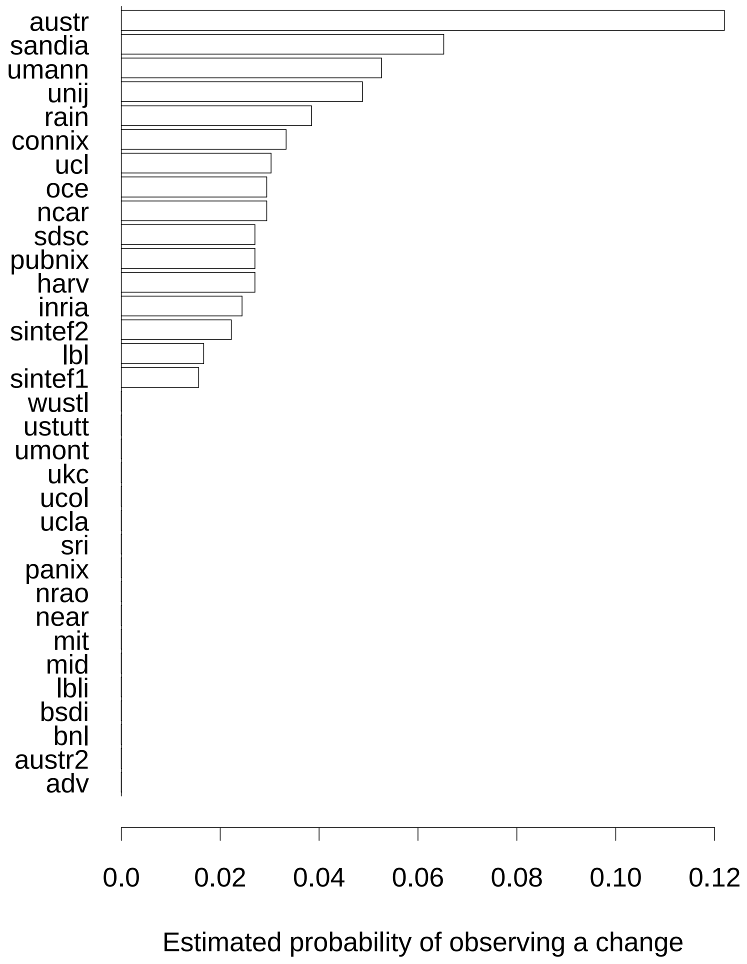

Figure 7.4: Site-to-site variation in $P^{10}_{\mathrm{dst\ s}}$

into Australia.[8] The changes are either the first Australian hop of `vic.gw.au`, in Melbourne, or `act.gw.au`, in Canberra, or `serial4-6.pad-core2.sydney.telstra.net` in Sydney followed by an additional hop to `nsw.gw.au` (also in Sydney). These are the only points of change: before and after, the routes are unchanged. Thus, the destination `austr` exhibits rapid (time scale of tens of minutes) changes in its incoming routing, and these changes are non-negligible, since they involve different Australian cities. As such, the routing *to* `austr` is not at all persistent.

However, for the next potential outlier, `sandia`, the story is different. Both of its changes occurred along the path originating at `sri`, and reflected the following change at hops 8 and 9:

```
core-fddi-0.sanfrancisco.mci.net
borderx2-fddi0-0.sanfrancisco.mci.
```

versus:

```
core2-fddi-0.sanfrancisco.mci.net
borderx2-fddi-1.sanfrancisco.mci.net
```

These changes are localized to a single city. Furthermore, had this change been more prevalent, we might have decided that the two pairs of routers in question were "tightly coupled" (§ 7.4), except that it turns out that they are responsible for routing changes only between `sri` and `sandia`. Thus, we can deal with this outlier by just eliminating the path `sri` $\Rightarrow$ `sandia`, but keeping the other paths with destination `sandia`.

In addition to the destination `austr`, a similar analysis of $P^{10}_{\text{src s}}$ points up `ucl`, `ukc`, `mid`, and `umann` as outliers. Both `ucl` and `ukc` had frequent oscillations in the routers visited between London and Washington, D.C., alternating between the two hops of:

```
icm-lon-1.icp.net
icm-dc-1-s3/2-1984k.icp.net
```

and the four hops of:

```
eu-gw.ja.net
gw.linx.ja.net
us-gw.thouse.ja.net
icm-dc-1-s2/4-1984k.icp.net
```

Note that these different hops also correspond to different AS's, as the latter includes AS 786 (JANET) and the former does not. For `mid` and `umann`, however, the changes did not have a clear pattern, and their prevalence could be due simply to chance.

On the basis of this analysis, we conclude that the sources `ucl` and `ukc`, and the destination `austr`, suffer from significant, high-frequency oscillation, and excluded them from further analysis. After removing any measurements originating from the first two or destined to `austr`, we then revisited the range of values for $P^{10}_{\text{src s}}$ and $P^{10}_{\text{dst s}}$. Both of these now had a median of 0 observed changes, and a maximum corresponding to about 1 change per hour (this latter rate is computed by dividing the number of route changes observed for the site's paths by total amount of time spanned by the measurements of those paths). On this basis, we believe we are on firm ground treating pairs of measurements between these sites, made less than an hour apart, both observing the same route, as consistent with that route having persisted unchanged between the measurements.

[8]Note that in general the paths to `austr` and `austr2` use two different trans-Pacific links, which is why `austr2` does not exhibit these rapid changes.

7.6.2 Medium-scale route alternation

Given the findings in the previous section that, except for a few sites, route changes do not occur on time scales less than an hour, we now turn to analyzing those measurements made an hour or less apart to determine what they tell us about medium-scale routing persistence. We proceed much as in § 7.6.1.

Let $P^{\text{hr}}_{\text{src s}}$ and $P^{\text{hr}}_{\text{dst s}}$ be the analogs of $P^{10}_{\text{src s}}$ and $P^{10}_{\text{dst s}}$, but now for measurements made an hour or less apart. After eliminating the rapidly oscillating paths identified in the previous section, we have 7,287 pairs of measurements to assess.

The data also included 1,517 triple observations spanning an hour or less. Of these, only 10 observed the pattern R_1, R_2, R_1 or R_1, R_2, R_3, indicating that, in general, two observations spaced an hour apart are not likely to miss a routing change.

Plots similar to Figure 7.4 immediately pick out paths originating from `bnl` as exhibiting rapid changes. These changes are almost all from oscillation between `llnl-satm.es.net` and `pppl-satm.es.net`. The first of these is in Livermore, California, while the other is in Princeton, New Jersey, so this change is definitely major. ESNET oscillations also occurred on one-hour time scales in traffic between `lbl` (and `lbli`) and the Cambridge sites, `near`, `harv`, and `mit`.

The other prevalent oscillation we found was between the source `umann` and the destinations `ucl` and `ukc`. Here the alternation was:

```
ch-s1-0.eurocore.bt.net
uk-s1-1.eurocore.bt.net
```

which goes through Switzerland to reach England, versus

```
nl-s1-1.eurocore.bt.net
uk-s1-0.eurocore.bt.net
```

which goes through the Netherlands instead, also a major change.

Eliminating these oscillating paths leaves us with 6,919 measurement pairs. These paths are not statistically identical (i.e., we find among them paths that have significantly different route change rates), but all have low rates of routing changes. For these paths, the median $P^{\text{hr}}_{\text{src s}}$ and $P^{\text{hr}}_{\text{dst s}}$ correspond to one routing change per 1.5 days, and the maximum to one change per 12 hours.

7.6.3 Large-scale route alternation

Given that, after removing the oscillating paths discussed in § 7.6.1 and § 7.6.2, we expect at most on the order of one route change per 12 hours, we now can analyze measurements less than 6 hours apart of the remaining paths to assess longer-term route changes. There were 15,171 such pairs of measurements. As 6 hours is significantly larger than the mean 2 hour sampling interval (§ 7.6), not surprisingly we find many triple measurements spanning less than 6 hours. But of the 10,660 triple measurements, only 75 included a route change of the form R_1, R_2, R_1 or R_1, R_2, R_3, indicating that, for the paths to which we have now narrowed our focus, we are still not missing many routing changes using measurements spaced up to 6 hours apart.

Employing the same analysis, we first identify `sintef1` and `sintef2` as outliers, both as source and as destination sites. The majority of their route changes turn out to be oscillations between two sets of routers. The first alternates between:

```
trd-gw2.uninett.no
```

in Trondheim, and

```
oslos-gw.uninett.no
trds-gw.uninett.no
```

(or the reverse of this, for paths originating at `sintef1` or `sintef2`), which includes an extra hop to Oslo. The second alternates between:

```
nord-gw.nordu.net
no-gw.nordu.net
```

(or the reverse), the first hop in Stockholm and the second in Trondheim, and

```
syd-gw.nordu.net
no-gw2.nordu.net
oslos-gw.uninett.no
trds-gw.uninett.no
```

which again adds a visit to Oslo (middle two hops).

Two other outliers at this level are traffic to or from `sdsc`, which alternates between two different pairs of CERFNET routers, all sited in San Diego, and traffic originating from `mid`, which alternates between two MIDNET routers, both in St. Louis.

Eliminating these paths leaves 11,174 measurements of the 712 remaining paths. The paths between the sites in these remaining measurements are quite stable, with a maximum transition rate for any site of about one change every two days, and a median rate of one change every four days.

7.6.4 Duration of long-lived routes

We will term the remaining measurements as corresponding to "long-lived" routes. For these, we might hazard to estimate the durations of the different routes as follows. We suppose that we are not completely missing any routing transitions (changes of the form R_1, R_2, R_1, where we only observe the first and last). We base this assumption on the overall low rate of routing changes. Then, for a sequence of measurements all observing the same route, we assume that the route's duration was at least the span of the measurements. So if the last observation was made two weeks after the first observation, we assume the route's duration was at least two weeks. Furthermore, if at time t_1 we observe route R_1, and then the next measurement at time t_2 observes route R_2, we make a "best guess" that route R_1 terminated and route R_2 began half way between these measurements, i.e., at time $\frac{t_1+t_2}{2}$.

For routes observed at the beginning (end) of our measurement period, but not spanning the entire measurement period, we assign a starting (ending) time as follows. If the next (previous) measurement also observed the route, then we estimate that the route persisted for at least that much time into the past (future). If the next (previous) measurement did *not* observe the route, then we take the lone observation of the route as its starting (ending) time. This rule will tend to underestimate routing durations, while the rule in the previous paragraph will tend to overestimate (due to occasionally missing a routing change), so these estimation errors will to some degree tend to cancel.

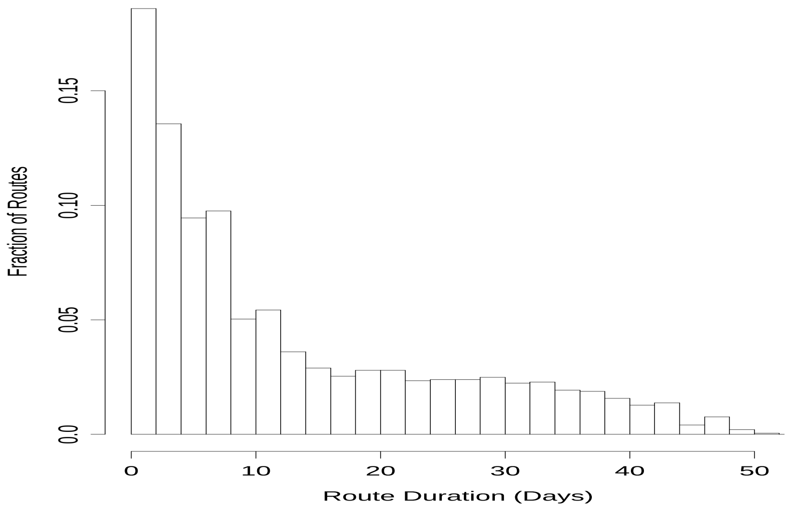

Figure 7.5: Estimated distribution of long-lived route durations

Figure 7.5 shows the distribution of the estimated durations of the "long-lived" routes. Even keeping in mind that our estimates are rough, it is clear that the distribution of long-lived route durations has two distinct regions, with many of the routes persisting for 1-7 days, and another group persisting for several weeks. (Although not evident from the plot, about 4% of the routes had durations under 6 hours, so we might consider the distribution as having three distinct regions.) About half the routes persisted for under a week, but the half of the routes lasting more than a week accounted for 90% of total persistence, meaning the integrated amount of time during which routes remained unchanged. This means that, if we observe a path at an arbitrary point in time, *and we are not observing one of the numerous, more rapidly oscillating paths outlined in the previous sections*, then we have about a 90% chance of observing a route for that path with a duration of at least a week.

7.6.5 Summary of routing persistence

We summarize routing persistence as follows. First, *routing changes occur over a wide range of time scales, ranging from seconds to days.* Table XII lists different time scales over which routes change. The second column gives the percentage of all of our measurement paths (source/destination pairs) that were affected by route changes at the given time scale. (The first two rows show "N/A" in this field because the changes were due to a very small set of routers, so we do not claim any sort of representative fractions.) The third column gives the section where we discuss the changes, and the final column any associated notes. When the note mentions "inside the network" or "intra-network," we mean that the changes occurred not at the stub networks where the sites themselves connect to the Internet, but instead in what we would deem the Internet infrastructure.

One important point apparent from the table is that routing changes on shorter time scales

Time scale	% Paths Affected	§	Notes
seconds	N/A	§ 6.6	"Flutter" for purposes of load balancing. Treated separately, as a pathology, and not included in the analysis of persistence.
minutes	N/A	§ 7.4	"Tightly-coupled routers." We identified five instances, which we merged into single routers for the remainder of the analysis.
10's of minutes	9%	§ 7.6.1	Frequent route changes inside the network. In some cases involved routing through different cities or AS's.
hours	4%	§ 7.6.2	Usually intra-network changes.
6+ hours	19%	§ 7.6.3	Also intra-network changes.
days	68%	§ 7.6.4	Two regions. 50% of routes persist for under 7 days. The remaining 50% account for 90% of the total route lifetimes.

Table XII: Summary of persistence at different time scales

(fewer than days) happen *inside the network* and not at the stub networks. Thus, *those changes observed in our measurements are likely to be similar to those observed by most Internet sites.*

On the other hand, while the changes occurred inside the network, only those involving `ucl` and `ukc` (§ 7.6.1) involved different sequences of autonomous systems. While this bodes well for the scalability of BGP, we do not claim this finding as having major significance: one could make a much more thorough assessment of the degree of inter-AS route flapping by analyzing the data discussed in [Do95, Me95b].

Finally, two thirds of the Internet paths we studied had quite stable paths, persisting for days or weeks. This finding is in accord with that of Chinoy's, who found that most networks are nearly quiescent (in terms of routing changes) while a few exhibit frequent connectivity transitions [Ch93].

7.7 Detecting route changes

Given our findings that routes change in the Internet on a wide range of time scales, we would like to find mechanisms by which an endpoint can detect that its route to a remote destination has changed. This knowledge has two different applications. The first is that it allows the endpoint to flush any cached information associated with the route, such as round-trip time or available bandwidth. The second application is for network measurement experiments. A number of Internet experiments have been made in which a path through the network is repeatedly sampled [Mi83, CPB93a, Bo93, SAGJ93, Mu94, BCG95]. For such measurements it is important to know whether each time the path is measured, the measurement is observing the same route for that path, or whether the route may have changed (affecting the measurement).

While `traceroute` can be used to elicit the route currently used for a given Internet path, its use is expensive in terms of network resources, and also slow because of the necessity to wait for (possibly dropped) replies to many probe packets.

Granularity	False positives	False negatives	Error rate
host	0%	25%	3%
city	4%	26%	5%
AS path	5%	10%	5%

Table XIII: Summary of TTL method for detecting route changes at different granularities

On the other hand, endpoints can easily determine whether a route's hop count has changed by seeing whether the TTL of packets arriving from the remote destination differ from the previously observed TTL. Because the IP TTL field is in fact a hop count and not a time-to-live (§ 4.2.1), this measurement has no noise, provided the remote destination always sends packets with the same initial TTL. Thus, the endpoint need receive only a single packet from the destination in order to detect that the hop count of the path from the destination to the endpoint has changed. We call this method the "TTL method." To our knowledge, it was first used in [CPB93a].

While the TTL method has an attractive simplicity, it will sometimes result in "false negatives": the underlying route might have changed, perhaps drastically, but if the new route happens to have the same number of hops as the cached one, the TTL method will report it as unchanged. In this section, we explore the degree to which these false negatives affect the practicality of the method.

After removing pathologies and fluttering paths, the data contained 30,145 consecutive `traceroutes` for us to test. Of these, 3,380 were route changes when viewed at host granularity, 1,928 at city granularity, and 1,266 at AS path granularity.

We consider a route to have changed if and only if it did not visit exactly the same hosts (cities; AS's) in the same order. Before determining the host visited at each hop, however, we merged the "tightly-coupled" routers discussed in § 7.4 into a single router.

We deem the method as generating a "false positive" if it erroneously declares that the route changed, and a "false negative" if it fails to detect that the route did indeed change. To make these notions more precise, suppose that, out of N observations, K were genuine route changes at a given granularity, but of these K the method only detects k, and it also erroneously "detects" b bogus route changes. Then the false positive rate is $b/(N-K)$, and the false negative rate is $(K-k)/K$. We can also define an overall "error rate," which is the proportion of time that the method misinforms us one way or the other: $(b+K-k)/N$.

Barring the remote host altering its initial TTL setting, or routers actually decrementing the TTL field for each second they delay a packet, the TTL method will never generate a false positive at host granularity[9]. It can do so at other granularities, however, when the underlying route changes in the number of hops, but the same cities or AS's are still visited. At all three granularities, the TTL method can generate false negatives.

Table XIII summarizes the effectiveness of the TTL method for detecting different granularities of route changes. Its overall error rate is consistently low. This is mostly a reflection of the fact that all-in-all the underlying route does not change very often. Because in the absence of any change whatsoever the TTL method always reports "no change," it is correct whenever the

[9]Provided we exclude from testing pathological routes that visit a given hop more than once, which we did.